Analysing Extrapolation Capabilities of Modern Positional Embeddings in Vision Transformers

Anonymous authors
Paper under double-blind review

Abstract

Vision Transformers scale quadratically with input resolution, making high-resolution training prohibitively expensive, yet detection and segmentation demand it. A natural alternative is resolution extrapolation: train at low resolution and deploy at higher resolution without fine-tuning. Whether this works depends entirely on the positional encoding. Modern encodings such as RoPE, ALiBi, YaRN, and FIRE enable striking length extrapolation in language models, but their behavior on 2D image grids is largely unknown. We present a systematic study, training a single ViT-Tiny on ImageNet-100 at 224×224 and evaluating zero-shot up to 1024×1024 ($4.57\times$) on classification, COCO detection, and ADE20K segmentation. Across all three tasks, additive attention-bias methods consistently outperform rotation-based methods at large scales: at $4.57\times$, FIRE retains 63.1% Top-1 accuracy while RoPE collapses to 18.9%. An attention-entropy analysis shows that bias-based encodings preserve focused, semantically coherent attention, whereas rotation-based phases drift out of distribution and induce attention collapse. These results recast extrapolation robustness as the primary axis for choosing positional encodings, and yield a practical recipe for resolution-flexible Vision Transformers.

1 Introduction

Vision Transformers (ViTs) Dosovitskiy et al. (2021) have emerged as a dominant architecture across computer vision, demonstrating strong performance in both recognition and dense prediction settings. However, ViTs suffer from quadratic computational complexity with respect to token count, making training at high resolutions prohibitively expensive. This limitation is particularly problematic for detection and segmentation, where fine-grained spatial detail is essential. Resolution extrapolation, i.e., training at low resolution and deploying at higher resolution, offers an appealing alternative, but is fundamentally constrained by the poor generalization of standard absolute positional embeddings beyond their training grid.

The practical importance of resolution extrapolation cannot be overstated. Object detection and semantic segmentation require precise localization of fine-grained structures, small objects, and thin boundaries, all of which benefit from higher-resolution input. Yet training a ViT-Base at 1024×1024 requires roughly $21\times$ the FLOPs compared to 224×224, making large-scale training at high resolution impractical for many research groups and deployment scenarios. If a model trained cheaply at low resolution could be deployed at higher resolution with minimal accuracy loss, it would dramatically reduce both the computational cost and the carbon footprint of modern vision systems. The critical bottleneck in realizing this goal is the positional encoding: while the feature extraction layers of a ViT are inherently resolution-agnostic (each patch is processed independently before attention), the positional encoding ties the model to a specific spatial grid and determines whether the learned spatial relationships generalize to unseen resolutions.

Recent progress in Large Language Models (LLMs) Touvron et al. (2023) has produced a rich landscape of positional encoding strategies specifically designed for length extrapolation. These methods fall into two broad families. *Rotation-based methods*, such as RoPE Su et al. (2024) and its extension YaRN Peng et al. (2024), encode relative positions by applying position-dependent complex rotations to query and key vectors, so that the attention score between two tokens depends only on their relative displacement. *Additive bias*

methods, such as ALiBi Press et al. (2022) and FIRE Li et al. (2024), instead inject positional information as a deterministic or learned bias added directly to attention logits, leaving the content-based query–key similarity untouched. Both families have enabled dramatic context-length extension in 1D language models, but their behavior on the two-dimensional spatial lattice of images, where relative offsets become 2D vectors and distances grow radially rather than linearly, remains largely unexplored.

A handful of prior works have adapted individual methods to vision, most notably 2D RoPE Heo et al. (2024), but no study has systematically compared rotation-based and bias-based families under controlled conditions across multiple downstream tasks. This gap is significant because the structural properties that make a positional encoding extrapolatable in 1D (translation invariance, monotonic attention decay, suppression of high-frequency oscillations) may not straightforwardly transfer to 2D grids, where the topology, distance metric, and boundary conditions all differ.

In this work, we conduct a systematic empirical study of modern positional encoding strategies adapted to 2D Vision Transformers. Due to computational constraints, we focus on ViT-Tiny trained on ImageNet-100 Deng et al. (2009) at 224×224, evaluating zero-shot extrapolation up to 1024×1024 ($4.57\times$ the training grid) without fine-tuning. We evaluate each method on classification, COCO object detection, and ADE20K semantic segmentation, and complement accuracy metrics with an attention-entropy analysis that reveals *why* certain methods fail. Our contributions are:

- We provide the first comprehensive comparison of RoPE Su et al. (2024); Heo et al. (2024), ALiBi Press et al. (2022), YaRN Peng et al. (2024), and FIRE Li et al. (2024) for 2D resolution extrapolation in Vision Transformers, evaluating across classification, detection, and segmentation.
- We demonstrate that attention bias-based methods (ALiBi, FIRE) substantially outperform rotation-based methods (RoPE, YaRN) under extrapolation, with FIRE retaining 63.1% accuracy at $4.6\times$ resolution versus RoPE's 18.9%.
- We analyze attention patterns to explain these differences, showing that bias-based methods maintain focused attention while rotation-based methods exhibit attention collapse at extrapolated resolutions.
- We identify an anomalous failure mode of ALiBi at very low resolutions (e.g., 6×6 patch grids), providing practical deployment guidance.

2 Related Works

2.1 Positional Encoding for Transformers

Transformers Vaswani et al. (2017) are permutation-invariant and require explicit positional information to model token order or spatial layout. Early methods employ absolute positional embeddings, either fixed sinusoidal or learned, added directly to token features. Vision Transformers Dosovitskiy et al. (2021) adopt learned 2D absolute embeddings defined on a fixed spatial grid, using bicubic interpolation Beyer et al. (2022) when transferring to higher resolutions. However, such embeddings are resolution-specific and do not provide true zero-shot extrapolation.

Relative positional embeddings (RPEs) instead encode pairwise displacements by modifying attention scores, ensuring translation invariance and enabling generalization to unseen positions. In NLP, methods such as those of Shaw *et al.* Shaw et al. (2018) and T5 Raffel et al. (2020) improve sequence-length generalization, while in vision, Swin Transformer Liu et al. (2021) introduces learnable 2D relative position biases within local windows. More recent approaches such as FlexiViT Beyer et al. (2023) and conditional positional encodings Chu et al. (2023) explore resizing strategies for both patch and positional embeddings, but still rely on bounded lookup tables or learned absolute embeddings with limited support for long-range spatial extrapolation.

The challenge of length extrapolation has been most actively studied in large language models. Several key properties have been identified for successful extrapolation Hong et al. (2024); Dong et al. (2024):

translation-invariant dependence on relative offsets, monotonic or smoothly decaying attention with distance, and suppression of high-frequency oscillations that destabilize long-range correlations. Building on these principles, a range of methods have been proposed, including RoPE Su et al. (2024), ALiBi Press et al. (2022), YaRN Peng et al. (2024), FIRE Li et al. (2024), position interpolation Chen et al. (2024), XPOS Sun et al. (2022), InfLLM Xiao et al. (2024), divide-and-conquer scaling Yang et al. (2025), mixed-radix extensions Tian et al. (2026), imaginary RoPE extensions Liu et al. (2025), multi-scale self-injection Han et al. (2026), and context window scheduling Zhu et al. (2025). While these mechanisms have enabled dramatic context extension in 1D, existing analyses are almost exclusively confined to linear token sequences; how these extrapolation principles transfer to 2D spatial grids, where offsets become vectors and distances grow radially, remains largely unexplored.

2.2 Positional Embeddings in Vision

In the visual domain, the adaptation of modern positional encodings is still in its early stages. The original ViT and subsequent works Dosovitskiy et al. (2021); Beyer et al. (2022) interpolate absolute positional embeddings at new resolutions, enabling fine-tuning but not zero-shot extrapolation. Among relative methods, 2D RoPE Heo et al. (2024) has been applied to vision transformers by extending rotary embeddings to two spatial axes, but has not been systematically compared against other modern encodings. ALiBi, FIRE, and YaRN, despite their proven 1D extrapolation capabilities, have received little attention in the vision literature.

This gap is our primary motivation. Existing studies evaluate individual methods in isolation or focus exclusively on classification; no prior work provides a controlled comparison of rotation-based and bias-based positional encodings across classification, detection, and segmentation under zero-shot resolution extrapolation. Our study fills this gap by benchmarking five representative methods under identical training and evaluation conditions.

3 Methodology

This section describes each positional encoding strategy evaluated in our study. We begin with the baseline Vision Transformer using learned absolute positional embeddings, then present the rotation-based and bias-based relative positional methods adapted for 2D spatial inputs.

3.1 Simple ViT: Learned Absolute Positional Embeddings

The standard Vision Transformer (ViT) Dosovitskiy et al. (2021); Beyer et al. (2022) processes an image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ by partitioning it into a grid of non-overlapping patches of size $P \times P$, yielding $N = HW/P^2$ patch tokens. Each patch is linearly projected to a d-dimensional embedding via a convolutional stem, and a learnable classification token $\mathbf{x}_{\text{cls}} \in \mathbb{R}^d$ is prepended to the sequence. Positional information is injected by adding a set of learnable absolute positional embeddings $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{N \times d}$ to the patch tokens:

$$\mathbf{z}_0 = [\mathbf{x}_{\text{cls}};\, \mathbf{x}_1\mathbf{E};\, \mathbf{x}_2\mathbf{E};\, \cdots ;\, \mathbf{x}_N\mathbf{E}] + \mathbf{E}_{\text{pos}}, \tag{1}$$

where $\mathbf{E} \in \mathbb{R}^{P^2C \times d}$ is the patch projection matrix. The resulting sequence is passed through L transformer encoder blocks, each consisting of multi-head self-attention (MSA) and an MLP with layer normalization and residual connections:

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \tag{2}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell. \tag{3}$$

The final representation of the classification token $\mathbf{z}_0^L$ is used for prediction.

Since the learned positional embeddings are defined on a fixed $H/P \times W/P$ grid, they cannot directly generalize to spatial grids of different sizes. When evaluating at a resolution different from training, we apply 2D bicubic interpolation to the positional embeddings, reshaping them to the original grid, interpolating to the target grid size, and flattening back to a sequence. While this enables inference at novel resolutions, the interpolated embeddings are only an approximation and may degrade as the resolution gap increases.

3.2 RoPE: Rotary Position Embeddings for 2D Vision

Rotary Position Embedding (RoPE) Su et al. (2024) encodes positional information by applying position-dependent complex rotations directly to the query and key vectors in self-attention, rather than adding an explicit positional embedding to the input tokens. For a token at 1D position n, RoPE defines a rotation matrix using frequency θ_t for the t-th dimension pair:

$$\mathbf{R}(n,t) = e^{i\theta_t n}, \quad \theta_t = \theta_{\text{base}}^{-t/(d_{\text{head}}/2)}, \tag{4}$$

where d_{head} is the per-head dimension and θ_{base} controls the frequency range. The rotation is applied to query and key vectors via the Hadamard product:

$$\bar{\mathbf{q}}'_n = \bar{\mathbf{q}}_n \circ \mathbf{R}(n), \quad \bar{\mathbf{k}}'_m = \bar{\mathbf{k}}_m \circ \mathbf{R}(m), \tag{5}$$

where $\bar{\mathbf{q}}, \bar{\mathbf{k}} \in \mathbb{C}^{d_{\text{head}}/2}$ are complex-valued views obtained by pairing consecutive dimensions. The resulting attention score between positions n and m naturally encodes their relative displacement:

$$A'_{(n,m)} = \text{Re}[\bar{\mathbf{q}}_n \bar{\mathbf{k}}^*_m \, e^{i\theta_t(n-m)}]. \tag{6}$$

To adapt RoPE to the 2D spatial structure of Vision Transformers, we follow the mixed learnable frequency formulation Heo et al. (2024). Each patch has a 2D position $\mathbf{p}_n = (p_n^x, p_n^y)$ on the spatial grid. The rotation matrix is defined using separate frequency vectors for each axis:

$$\mathbf{R}(n,t) = e^{i(\theta_t^x \, p_n^x + \theta_t^y \, p_n^y)}, \tag{7}$$

where (θ_t^x, θ_t^y) are per-head, per-layer learnable frequency parameters. Unlike axial RoPE, which restricts each frequency dimension to a single spatial axis, the mixed formulation allows each frequency to attend to both horizontal and vertical directions simultaneously. This enables RoPE to capture diagonal spatial relationships, which axial frequencies cannot represent. The resulting attention matrix encodes relative 2D displacements:

$$A'_{(n,m)} = \text{Re}\Big[\bar{\mathbf{q}}_n \bar{\mathbf{k}}^*_m \, e^{i(\theta_t^x (p_n^x - p_m^x) + \theta_t^y (p_n^y - p_m^y))}\Big]. \tag{8}$$

In our implementation, the mixed frequencies are initialized from a base frequency schedule with random per-head axis rotations and are trained jointly with the network parameters. Rotary embeddings are applied to all patch tokens but skipped for the classification token, which has no spatial position. For detection and segmentation backbones that omit the classification token, RoPE is applied to all tokens directly. Since RoPE does not introduce fixed-size positional parameters, it naturally handles variable-length sequences; however, at resolutions beyond the training range, the effective phase angles move outside the distribution seen during training, which can degrade attention stability.

3.3 FIRE: Functional Interpolation for Relative Positional Encoding

FIRE (Functional Interpolation for Relative Positional Encoding) Li et al. (2024) models relative positional information as a continuous, learnable attention bias designed for length extrapolation. Instead of relying on discrete lookup tables, FIRE parameterizes the attention bias between a query position i and a key position j using a small MLP:

$$b(i,j) = f_\theta\left(\frac{\psi(|i-j|)}{\psi(\max\{L,i\})}\right), \tag{9}$$

where $f_\theta : \mathbb{R} \to \mathbb{R}^H$ outputs head-wise biases, $\psi(x) = \log(1 + cx)$ is a monotonic concave transform with learnable scale c, and L is a reference length controlling the onset of extrapolation. The normalization by $\max\{L, i\}$ implements *progressive interpolation*, constraining the MLP input to $[0, 1]$ for arbitrary sequence lengths and ensuring that longer contexts correspond to interpolation rather than out-of-distribution extrapolation in function space.

To adapt FIRE to Vision Transformers, we apply the formulation on the flattened 2D patch sequence. For an image tokenized into an $H \times W$ grid with $N = HW$ patches, we index patches by their flattened order

$i \in \{0, \ldots, N-1\}$ and compute pairwise relative distances $|i-j|$. The resulting bias matrix $\mathbf{B} \in \mathbb{R}^{H \times N \times N}$ is shared across layers and added directly to the scaled dot-product attention:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}} + \mathbf{B}\right) V. \tag{10}$$

Following the FIRE design, we incorporate (i) logarithmic distance warping to allocate higher resolution to local spatial relationships, (ii) thresholded normalization with a learnable multiplier on L to preserve short-range behavior, and (iii) a learnable global bias scale to stabilize optimization when transferring from 1D sequences to dense 2D grids. For classification models, the bias is padded so that the class token receives zero positional offset, while for dense prediction backbones the formulation operates directly on patch tokens without a class token.

This continuous, resolution-agnostic attention bias eliminates the need for positional embedding interpolation or frequency rescaling and naturally generalizes to larger patch grids. As a result, FIRE provides a principled mechanism for zero-shot spatial extrapolation in Vision Transformers, aligning closely with the functional interpolation theory of the original formulation while remaining fully compatible with standard ViT and downstream detection and segmentation backbones.

3.4 YaRN: Yet another RoPE extension for Resolution Extrapolation

YaRN (Yet another RoPE eNhancement) Peng et al. (2024) is a frequency-aware extension of Rotary Position Embeddings designed to enable robust context-length extrapolation by selectively rescaling rotational frequencies and attention magnitudes. Unlike uniform interpolation of RoPE Chen et al. (2024), YaRN preserves high-frequency components responsible for local ordering while only interpolating low-frequency dimensions that encode global structure.

Recall that RoPE applies a complex rotation to query and key vectors:

$$\tilde{\mathbf{q}}_i = \mathbf{q}_i \odot e^{\mathrm{i}\theta(i)}, \qquad \tilde{\mathbf{k}}_j = \mathbf{k}_j \odot e^{\mathrm{i}\theta(j)}, \tag{11}$$

where $\theta_d(i) = i/\lambda_d$ and λ_d denotes the wavelength of the d-th frequency. Position interpolation methods scale all dimensions uniformly, which compresses local relative angles and degrades short-range discrimination. YaRN instead performs *targeted interpolation* by modifying the RoPE frequencies as

$$\theta'_d = (1 - \gamma_d)\frac{\theta_d}{s} + \gamma_d \theta_d, \tag{12}$$

where s is the extrapolation factor and $\gamma_d \in [0, 1]$ is a ramp function determined by the ratio between the original context length and the wavelength λ_d. Low-frequency dimensions (global structure) are interpolated, while high-frequency dimensions (local structure) are preserved.

In addition, YaRN introduces attention temperature scaling by rescaling the rotary embeddings, effectively modifying the attention logits as

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{t\sqrt{d}}\right) V, \tag{13}$$

where t is a scale-dependent temperature that stabilizes attention entropy under long-context extrapolation.

For Vision Transformers, we adapt YaRN to the 2D spatial domain by applying RoPE independently along horizontal and vertical axes. Let (x_i, y_i) denote the coordinates of the i-th patch. The complex rotation is computed using mixed 2D frequencies:

$$\theta(i) = \omega_x x_i + \omega_y y_i, \tag{14}$$

with YaRN-scaled frequency matrices ω_x, ω_y constructed per attention head and per layer. In our implementation, the modified frequencies are precomputed using the NTK-by-parts ramp and attention scaling, and the resulting complex exponentials are applied to queries and keys before attention computation.

This 2D YaRN formulation yields resolution-agnostic rotary embeddings that preserve local spatial precision while enabling stable zero-shot extrapolation to significantly larger patch grids than those observed during training.

Table 1: Comparison of positional encoding methods. "Pos. Params" indicates learnable positional parameters; "Res. Agnostic" indicates whether the method handles arbitrary resolutions without interpolation.

Method	**Type**	**Pos. Params**	**Res. Agnostic**
Simple ViT	Absolute	$N \times d$	No (interp.)
RoPE	Rotation	Freq. vectors	Partial
YaRN	Rotation	Same as RoPE	Yes (scaling)
ALiBi	Add. bias	0	Yes
FIRE	Add. bias	MLP ($\sim$1K)	Yes

3.5 ALiBi: Attention with Linear Biases for Vision Transformers

ALiBi Press et al. (2022) incorporates relative positional information into self-attention by adding a deterministic, head-specific linear bias directly to the attention logits. Unlike absolute positional embeddings, ALiBi does not encode positions explicitly, enabling robust extrapolation to longer sequences and higher resolutions without interpolation or additional learned positional parameters.

For a sequence indexed by positions i and j, the attention bias for head h is defined as

$$B_{ij}^{(h)} = -m_h \cdot d(i,j),$$

where m_h is a fixed slope assigned to each head and $d(i,j)$ denotes the relative distance between tokens. The slopes decrease monotonically across heads, allowing different heads to focus on different context ranges. Since these slopes are fixed, ALiBi introduces no trainable parameters for positional encoding.

In vision transformers, tokens correspond to image patches arranged on a two-dimensional grid with spatial coordinates (x_i, y_i). The relative distance between two patches is computed using the Euclidean metric

$$d(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2},$$

preserving spatial inductive bias while remaining resolution-agnostic.

The classification token, which lacks a spatial location, is assigned a constant distance to all patch tokens, resulting in a uniform CLS-to-patch bias while maintaining meaningful patch-to-patch relationships. The resulting bias tensor is shared across all transformer layers.

During attention computation, the ALiBi bias is additively incorporated into the scaled dot-product attention,

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + B\right)V,$$

directly modulating attention scores based on relative distance. The model otherwise follows a standard Vision Transformer architecture, with patch embeddings generated by a convolutional stem and final predictions obtained from the encoded CLS token.

3.6 Summary of Positional Encoding Methods

Table 1 summarizes the key properties of each positional encoding method. A fundamental distinction is between *additive bias* methods (ALiBi, FIRE), which inject positional information directly into attention logits, and *rotation-based* methods (RoPE, YaRN), which encode position through query-key rotations. Bias-based methods are inherently resolution-agnostic since they compute distances on-the-fly, while rotation-based methods require careful frequency design to avoid out-of-distribution phases at extrapolated resolutions.

4 Experiments

4.1 Experimental Setup

Architecture. Due to computational constraints, we conduct all experiments using a ViT-Tiny backbone with hidden dimension $d = 192$, depth $L = 12$, $H = 3$ attention heads, MLP dimension 768, and patch size $P = 16$. This yields approximately 5.7M parameters across all variants and ensures that any performance differences are attributable solely to the positional encoding mechanism. While larger models may exhibit different absolute performance, we expect the relative trends between positional encoding methods to generalize, as confirmed by the consistency of our findings across three diverse tasks.

Training: Classification. Models are trained from scratch on ImageNet-100 (a 100-class subset of ImageNet-1K Deng et al. (2009)) at resolution 224×224 for 300 epochs. We use AdamW Loshchilov & Hutter (2019) with base learning rate 10^{-3}, minimum learning rate 10^{-5}, weight decay 0.05, and $(\beta_1, \beta_2) = (0.9, 0.999)$. The learning rate follows a cosine schedule with 10 epochs of linear warmup from 10^{-6}. We apply label smoothing with factor 0.1, gradient clipping at norm 1.0, and automatic mixed precision. Data augmentation includes random resized cropping and horizontal flipping. For models with learnable frequency parameters (RoPE and YaRN), frequency tensors and bias/normalization parameters are excluded from weight decay.

Training: Detection. For object detection, we fine-tune on COCO 2017 Lin et al. (2014) using Faster R-CNN Ren et al. (2015) with an FPN Lin et al. (2017) neck built on top of each ViT-Tiny backbone. Models are initialized from the corresponding ImageNet-100 classification checkpoint. Training follows standard MMDetection Chen et al. (2019) protocols at a base resolution of 512×512.

Training: Segmentation. For semantic segmentation, we fine-tune on ADE20K Zhou et al. (2017) using UPerNet Xiao et al. (2018) with each ViT-Tiny backbone. Models are initialized from classification checkpoints and trained following standard MMSegmentation MMSegmentation Contributors (2020) protocols at a base resolution of 512×512.

Positional Embedding Variants. We compare the following methods:

- **Simple ViT**: Learned absolute 2D positional embeddings with bicubic interpolation at test time.
- **RoPE ViT**: 2D Rotary Position Embeddings with per-head random axis rotations ($\theta = 10$).
- **YaRN ViT**: Trained identically to RoPE (loaded from the same checkpoint); at inference, NTK-by-parts frequency scaling is applied dynamically based on the resolution ratio ($\theta = 10{,}000$, $\beta_{\text{fast}} = 32$, $\beta_{\text{slow}} = 1$).
- **ALiBi ViT**: Attention with Linear Biases using 2D Euclidean distance and fixed geometric head slopes. No learnable positional parameters.
- **FIRE ViT**: Continuous learnable attention bias parameterized by a small MLP with logarithmic distance warping and progressive interpolation.

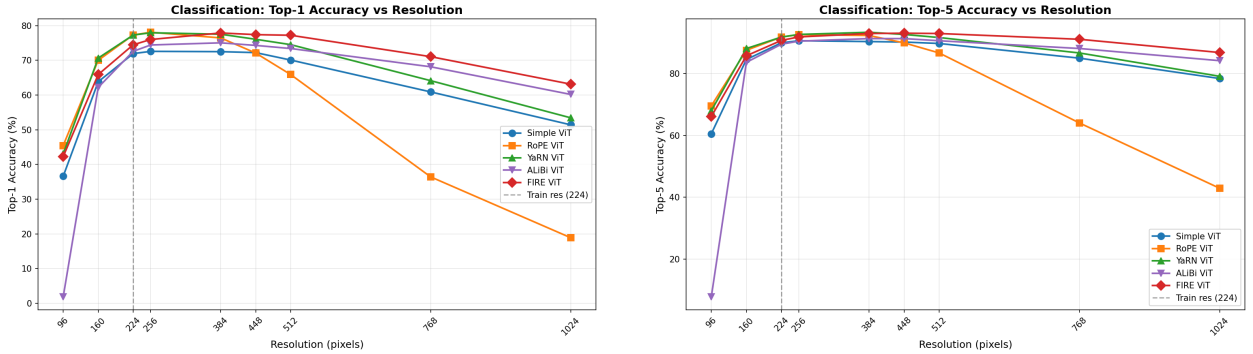
Evaluation Protocol. For classification, we evaluate each model on the ImageNet-100 validation set at nine resolutions: 96, 160, 224 (train), 256, 384, 448, 512, 768, and 1024 pixels. This corresponds to scale factors ranging from 0.43× (interpolation) to 4.57× (extreme extrapolation) relative to the 14×14 training patch grid. All evaluations are performed *zero-shot* without any fine-tuning at the target resolution. For detection and segmentation, we test at resolutions 384, 512 (train), 640, 768, and 1024 pixels.

4.2 Classification Results

Table 2 reports Top-1 and Top-5 accuracy across resolutions. All models peak at or slightly above the training resolution, but diverge significantly under extrapolation.

Table 2: Zero-shot classification accuracy (%) on ImageNet-100 at varying resolutions. All models are trained at 224×224. Bold indicates best per resolution; underline indicates second best. YaRN uses the same checkpoint as RoPE but applies NTK-by-parts frequency scaling at inference.

Model	Metric	Interpolation			Extrapolation					
		96 (0.43×)	**160** (0.71×)	**224** (1.0×)	**256** (1.14×)	**384** (1.71×)	**448** (2.0×)	**512** (2.29×)	**768** (3.43×)	**1024** (4.57×)
Simple ViT	Top-1	36.62	63.82	71.94	72.58	72.50	72.22	70.08	60.90	51.42
	Top-5	60.54	84.54	89.94	90.60	90.34	90.18	89.72	84.98	78.40
RoPE ViT	Top-1	**45.46**	70.00	**77.30**	**78.08**	76.46	72.10	65.90	36.42	18.90
	Top-5	**69.58**	87.52	91.80	92.60	92.30	89.92	86.66	64.04	42.94
YaRN ViT	Top-1	43.04	**70.64**	77.28	77.94	77.52	76.06	74.52	64.14	53.44
	Top-5	68.00	**88.08**	**91.80**	**92.64**	**93.34**	92.64	91.64	86.66	79.12
ALiBi ViT	Top-1	1.88	62.20	72.60	74.42	75.04	74.30	73.40	68.14	60.18
	Top-5	7.86	83.62	89.48	90.48	91.30	91.30	90.60	88.10	84.16
FIRE ViT	Top-1	42.28	65.90	74.48	76.00	**77.88**	**77.40**	**77.22**	**71.06**	**63.14**
	Top-5	66.10	85.74	90.72	91.84	92.94	**93.04**	**92.94**	**91.08**	**86.80**



(a) Top-1 Accuracy vs. Resolution

(b) Top-5 Accuracy vs. Resolution

Figure 1: Zero-shot classification extrapolation on ImageNet-100. All models are trained at 224×224 (dashed line) and evaluated up to 1024×1024 without fine-tuning. FIRE ViT and ALiBi ViT exhibit the most robust extrapolation behavior, while raw RoPE ViT degrades sharply beyond the training resolution.

Figure 1 visualizes these trends. While RoPE ViT attains the highest accuracy near the training resolution, FIRE ViT and ALiBi ViT remain substantially more stable under extreme resolution scaling.

Key Observations. **(i) Bias-based methods extrapolate best.** FIRE ViT achieves the highest accuracy at every resolution above 384, retaining 63.14% at 1024 (only 11.34 pp below its 224 baseline). ALiBi ViT exhibits a similarly graceful degradation profile, dropping just 12.42 pp from 224 to 1024. Both methods model relative distance through additive attention biases that are inherently resolution-agnostic.

(ii) Raw RoPE collapses at high extrapolation. RoPE ViT achieves the best accuracy at the training resolution (77.30%) but degrades catastrophically under extrapolation, falling to 18.90% at 1024, a 58.40 pp drop. This is consistent with the known sensitivity of RoPE to out-of-distribution positional frequencies.

(iii) YaRN significantly improves over raw RoPE. YaRN's NTK-by-parts scaling yields substantial gains over raw RoPE across all extrapolation regimes. At moderate extrapolation (77.52% vs. 76.46% at 384), the improvement is modest, but at extreme scales the gap widens dramatically: YaRN retains 53.44% at 1024 compared to RoPE's 18.90%, a 34.54 pp advantage. This demonstrates that selective frequency

Table 3: Detection performance (%) on COCO at varying resolutions. Trained at 512×512. AP_{50} provides a more interpretable view of the extrapolation trends.

Model	Metric	384	512	640	768	1024
Simple ViT	mAP	3.7	5.6	3.5	3.3	2.5
	AP_{50}	8.4	12.8	8.0	7.3	5.5
RoPE ViT	mAP	6.0	6.8	6.0	4.5	1.6
	AP_{50}	13.2	15.1	13.7	10.4	3.7
YaRN ViT	mAP	6.0	6.8	6.1	4.8	2.7
	AP_{50}	13.0	15.1	13.9	11.2	6.2
ALiBi ViT	mAP	3.0	3.4	3.4	3.3	2.7
	AP_{50}	6.6	7.6	7.6	7.2	5.9
FIRE ViT	mAP	3.4	3.4	3.2	2.9	2.3
	AP_{50}	7.4	7.5	6.8	6.1	4.7

scaling is highly effective for rotation-based methods, though YaRN still falls short of bias-based methods (FIRE: 63.14%, ALiBi: 60.18%) at the largest scale.

(iv) Simple ViT is surprisingly competitive. Bicubic interpolation of learned absolute embeddings provides a strong baseline, outperforming raw RoPE and YaRN at high extrapolation while remaining worse than ALiBi and FIRE.

4.3 Detection Results

Table 3 reports zero-shot bbox mAP on COCO under resolution extrapolation. Detection models are trained at 512×512 and evaluated at resolutions up to 1024.

In detection, RoPE ViT achieves the highest mAP at the training resolution (6.8%) but experiences the steepest decline, dropping to 1.6% at 1024, a 75% relative drop in AP_{50} (from 15.1% to 3.7%). ALiBi ViT demonstrates the most stable extrapolation profile, with only a 22% relative drop in AP_{50} from 512 to 1024. Notably, the overall mAP values are low due to the limited capacity of ViT-Tiny and training on ImageNet-100 rather than full ImageNet-1K; however, the *relative* degradation patterns are consistent with the classification findings and confirm that positional encoding robustness transfers to dense prediction.

Object Size Breakdown. Figure 2 decomposes detection performance by object scale (small, medium, large). A striking finding is that *large objects degrade most rapidly* under extrapolation across all methods. At 1024 px, mAP_L collapses to near-zero for most methods (< 0.01), while mAP_S for small objects actually *increases* for bias-based methods: ALiBi reaches 5.0% and FIRE reaches 4.5%. This counterintuitive result suggests that higher resolution provides more pixels per small object, improving detection, while large objects suffer from attention pattern fragmentation. RoPE exhibits the most severe large-object collapse ($9.0\% \rightarrow 0.1\%$), whereas ALiBi maintains the most balanced performance across scales.

4.4 Segmentation Results

Table 4 reports mIoU on ADE20K under resolution extrapolation.

Segmentation results reinforce the classification trends. ALiBi ViT retains 20.5% mIoU at 1024 (only 3.0 pp below its peak), while RoPE ViT degrades from 25.0% to 9.2%, a 15.8 pp collapse. FIRE ViT similarly maintains strong performance, dropping only 4.6 pp from peak to 1024. Simple ViT suffers a 10.3 pp decline, confirming that interpolated absolute embeddings provide weaker spatial coherence than relative or bias-based alternatives for dense prediction at novel resolutions.

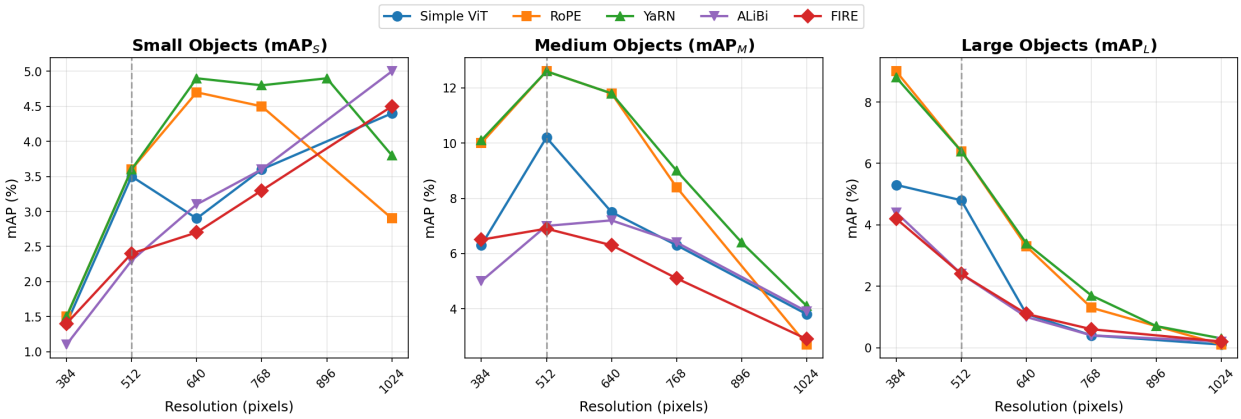


Figure 2: Detection mAP breakdown by object size across resolutions. Large objects (mAP_L) degrade most rapidly under extrapolation, while small object detection (mAP_S) improves at higher resolutions for bias-based methods.

Table 4: Segmentation mIoU (%) on ADE20K at varying resolutions. Trained at 512×512.

Model	384	512	640	768	1024
Simple ViT	16.5	21.8	16.0	14.5	11.5
RoPE ViT	<u>24.5</u>	**25.0**	23.1	19.3	9.2
YaRN ViT	**24.6**	<u>24.9</u>	<u>23.7</u>	<u>20.9</u>	15.4
ALiBi ViT	22.2	23.5	23.6	**23.0**	**20.5**
FIRE ViT	23.4	24.4	**24.2**	22.5	<u>19.8</u>

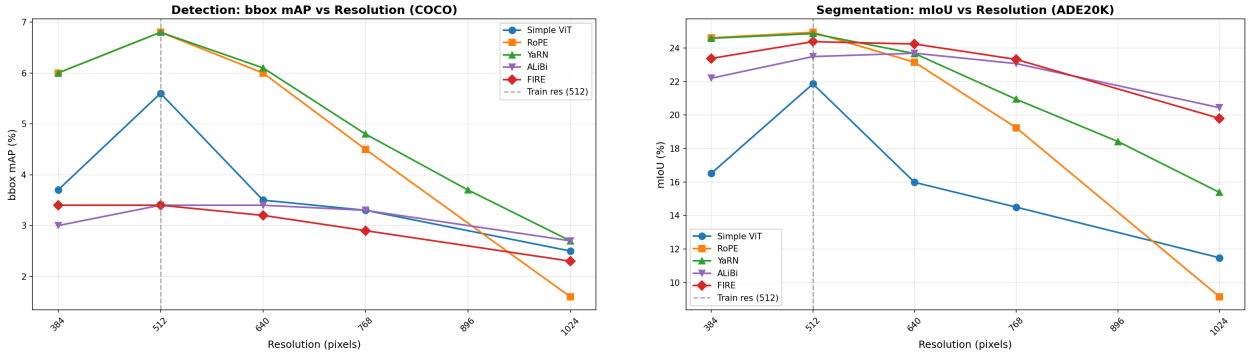
Figure 3 shows that the extrapolation trends observed in classification transfer to dense prediction tasks as well. Relative-bias methods remain substantially more stable than rotation-based methods as evaluation resolution increases.

4.5 Analysis

Bias vs. Rotation Methods. Across all three tasks, the additive bias methods (ALiBi and FIRE) consistently outperform the multiplicative rotation methods (RoPE and YaRN) under resolution extrapolation. We hypothesize that this is because additive biases directly modulate attention logits as a smooth function of distance, making them naturally resolution-agnostic. In contrast, rotation-based methods encode position through phase angles that grow linearly with coordinate magnitude; when extrapolated beyond the training range, these phases become out-of-distribution, causing destructive interference in the attention pattern.

The Role of Frequency Awareness. Among rotation-based methods, YaRN improves over raw RoPE through selective frequency scaling: it preserves high-frequency (local) components while interpolating low-frequency (global) ones via the NTK-by-parts ramp. While this yields gains at moderate extrapolation, YaRN still degrades substantially at extreme scales (4.57×), suggesting that the ramp parameters (β_{fast}, β_{slow}) may require resolution-specific tuning for 2D spatial data.

Consistency Across Tasks. The extrapolation rankings observed in classification (FIRE > ALiBi > Simple > YaRN > RoPE) are largely preserved in detection and segmentation, indicating that positional encoding robustness is a fundamental architectural property rather than a task-specific artifact. This consistency provides practical guidance: if a model extrapolates well in classification, it will likely do so in dense prediction as well.



(a) COCO bbox mAP vs. Resolution

(b) ADE20K mIoU vs. Resolution

Figure 3: Zero-shot extrapolation on dense prediction tasks. All models are trained at 512×512 (dashed line) and evaluated up to 1024×1024 without fine-tuning. ALiBi ViT and FIRE ViT maintain stable performance at higher resolutions, while RoPE ViT degrades sharply. YaRN provides moderate improvement over RoPE but still degrades at extreme scales.

5 Comparison and Analysis

This section synthesizes the empirical trends observed in Section 4 and provides a comparative interpretation of how different positional encoding mechanisms behave under resolution extrapolation. We focus on four recurring themes: (i) the distinct extrapolation profiles visible in classification, (ii) the structural difference between additive bias methods and rotation-based methods, (iii) the contrasting behavior of inference-time frequency scaling strategies, and (iv) the transfer of these trends to dense prediction tasks.

5.1 Classification Extrapolation Trends

The classification results reveal three qualitatively distinct operating regimes as the input resolution moves away from the training point. In the sub-training regime (96–160 px), all methods experience some degradation due to reduced spatial detail, although the extent varies substantially. In the near-training regime (224–384 px), several methods improve beyond their training-resolution accuracy, indicating that higher-resolution inputs can provide genuinely useful additional spatial information when the positional encoding remains well behaved. In the far-extrapolation regime (448–1024 px), the choice of positional encoding becomes the dominant factor, and the performance gap between methods widens sharply.

Interpolation Regime Anomaly. Figure 4 reveals a surprising finding in the interpolation regime: ALiBi ViT *collapses* to 1.88% Top-1 accuracy at 96 px, just 2.6% of its training-resolution performance, while all other methods retain 36–45% accuracy. This asymmetric failure is unique to ALiBi. We hypothesize that at very low token counts ($6 \times 6 = 36$ patches at 96 px), ALiBi's distance-based attention bias becomes *too strong relative to content-based similarity*, effectively overwhelming the learned representations. The bias magnitude scales with distance but the total spatial extent shrinks, causing disproportionate penalties. At 160 px ($10 \times 10 = 100$ patches), ALiBi recovers to 62.2%, indicating that the failure mode is specific to very small spatial grids. This finding has practical implications: ALiBi may require recalibration of slope parameters when deployed at resolutions significantly below the training resolution.

At the largest evaluated scale (1024 px, corresponding to 4.57× the training grid), the Top-1 ranking is: FIRE (63.14%) > ALiBi (60.18%) > YaRN (53.44%) > Simple ViT (51.42%) > RoPE (18.90%). The gap between the best and worst methods at this scale is 44.24 percentage points, which is substantially larger than the 5.36-point spread at the training resolution (71.94%–77.30%). This demonstrates that extrapolation robustness, rather than in-distribution accuracy alone, is the primary differentiator among positional encoding strategies in resolution-flexible Vision Transformers.

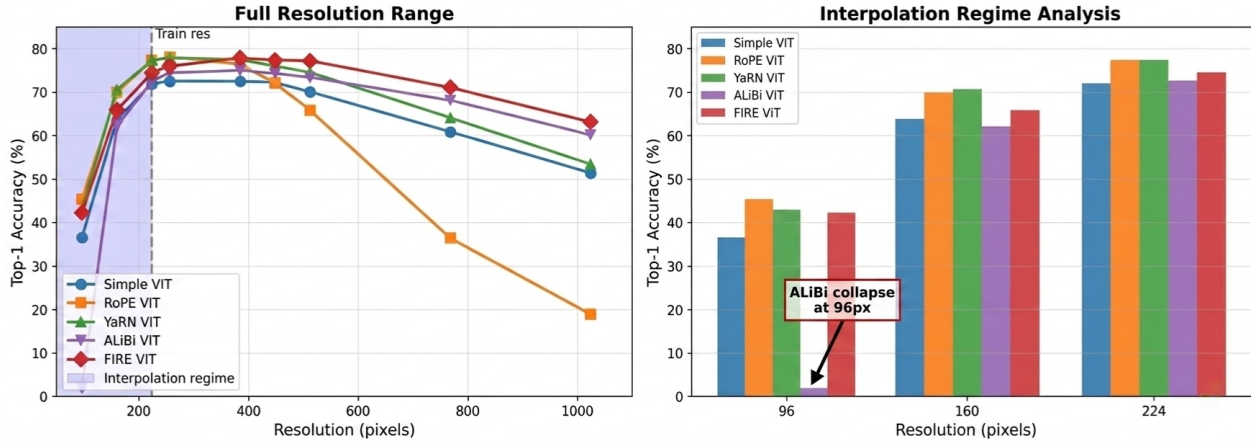


Figure 4: Interpolation regime analysis. Left: full resolution sweep with interpolation regime highlighted. Right: bar chart of sub-training resolutions showing ALiBi's anomalous collapse at 96 px (1.88%) while other methods retain 36–45% accuracy.

The Top-5 curves reinforce the same ordering while showing a slightly more compressed spread. At 1024 px, FIRE achieves the highest Top-5 accuracy (86.80%), followed by ALiBi (84.16%) and YaRN (79.12%), while vanilla RoPE drops to 42.94%. This indicates that the strongest methods degrade gracefully: even when the top prediction changes, the correct class often remains within the top few candidates, suggesting that the underlying representation remains semantically coherent over a broad range of resolutions. Notably, YaRN's NTK-by-parts scaling substantially improves its Top-5 retention compared to raw RoPE, nearly doubling it at 1024 px.

5.2 Bias-Based vs. Rotation-Based Methods

A consistent pattern across the experiments is the stronger extrapolation behavior of additive bias-based methods (ALiBi and FIRE) compared to rotation-based methods (RoPE and YaRN). The key architectural distinction is that bias-based methods inject positional information as an additive term in the attention logits, whereas rotation-based methods modify the query and key representations themselves before the dot product is computed.

This difference has important consequences under extrapolation. In bias-based methods, positional information acts as a spatial prior that encourages locality or relative structure, but the underlying content-based similarity remains intact. As the resolution increases, the positional bias simply extends to larger spatial distances, which generally leads to gradual degradation rather than abrupt failure. This behavior is clearly visible in both ALiBi and FIRE, which remain strong even at 1024 px. FIRE is the most robust overall in classification, dropping only 11.34 percentage points from its 224 px Top-1 accuracy (74.48%) to 1024 px (63.14%), while ALiBi drops only 12.42 points over the same range.

In contrast, rotation-based methods encode position directly into the geometry of the query–key interaction. This yields strong in-distribution performance: RoPE achieves the highest Top-1 accuracy at the training resolution (77.30%) and the peak score at 256 px (78.08%), but becomes increasingly fragile when evaluated outside the training range. As the patch grid grows, the effective positional phases move beyond the range seen during training, and the resulting attention patterns become progressively less stable. This is most visible in vanilla RoPE, which falls from 77.30% at 224 px to 18.90% at 1024 px, a 58.40-point collapse.

A notable secondary result is the strength of the Simple ViT baseline. Despite relying on learned absolute embeddings with bicubic interpolation, it reaches 51.42% Top-1 at 1024 px, remaining competitive with YaRN (53.44%) under extreme extrapolation. This suggests that smooth interpolation of learned absolute embeddings provides a surprisingly strong "safety floor" when more sophisticated relative schemes become unstable outside their training range. However, YaRN's selective frequency scaling enables it to outperform

Simple ViT by maintaining better local spatial precision while still benefiting from interpolated global structure.

5.3 Attention Pattern Analysis

To provide direct visual evidence of extrapolation failure, we examine attention maps from the final transformer layer across resolutions. Figure 5 shows CLS token attention overlaid on a sample image at 224, 512, and 768 px for each positional encoding method.

At the training resolution (224 px), all methods produce semantically meaningful attention patterns, focusing on the bird's head and body with entropy values around 900–1000. The differences become stark at 512 px: ALiBi's entropy *decreases* to 883.87, indicating that its attention becomes *more concentrated* as resolution increases, likely because the finer spatial granularity allows more precise localization. In contrast, RoPE's entropy jumps to 6928.47 (7× increase), and Simple ViT reaches 6966.18. Visually, ALiBi's attention remains tightly focused on the bird, while RoPE and Simple ViT begin attending to uninformative background regions (sky).

At 768 px, all methods show elevated entropy (16000–17500), reflecting the fundamental challenge of extrapolating to 3.4× the training grid size. However, qualitative differences persist: ALiBi (16266.66) and FIRE (16357.67) still exhibit visible concentration on the subject, whereas RoPE (17472.26) and Simple ViT (17501.27) produce nearly uniform attention distributions. This explains why bias-based methods retain reasonable classification accuracy at extreme scales: even when attention spreads, it maintains *relative* focus on semantically relevant regions, whereas rotation-based methods lose this spatial coherence entirely.

These attention patterns directly explain the accuracy trends observed in Section 4: the graceful degradation of ALiBi and FIRE corresponds to gradually diffusing but still object-centered attention, while RoPE's sharp accuracy collapse corresponds to attention that becomes spatially incoherent, effectively treating all image regions as equally relevant.

5.4 Dense Prediction and Practical Implications

The dense prediction experiments broadly support the trends observed in classification, but they also reveal an important nuance: the absolute ranking is not identical across all tasks, particularly in detection. In object detection, RoPE achieves the highest mAP near the training regime, reaching 6.8% at 512 px and remaining strong at 640 px (6.0%), but it also exhibits the steepest degradation, falling to 1.5% at 1024 px. YaRN, applying the same checkpoint with NTK-by-parts scaling, matches RoPE at 512 px (6.8%) but degrades more gracefully to 2.7% at 1024 px, nearly double RoPE's retained performance. By contrast, ALiBi shows the most stable detection profile, decreasing only from 3.4% at 512 px to 2.7% at 1024 px. FIRE remains competitive and relatively stable, dropping from 3.4% to 2.3%, but under the current ViT-Tiny setting it does not surpass RoPE in peak detection mAP.

In semantic segmentation, the extrapolation pattern aligns more closely with the classification results. RoPE achieves the highest score at the training resolution (25.0% mIoU at 512 px), but degrades sharply to 9.2% at 1024 px. YaRN significantly mitigates this degradation, retaining 15.4% mIoU at 1024 px compared to RoPE's 9.2%, a 6.2 pp improvement from the same checkpoint. By contrast, ALiBi retains 20.5% at 1024 px, and FIRE retains 19.8%, indicating substantially stronger spatial robustness at large resolutions. FIRE also exhibits the strongest peak among the stable methods, reaching 24.4% at 512 px and remaining above 24% at 640 px, while ALiBi maintains a flatter and more consistent curve across the full resolution sweep. YaRN occupies a middle ground, preserving more spatial coherence than raw RoPE while falling short of the bias-based methods.

These results suggest that positional encoding failures are amplified in dense prediction, where the model must maintain spatially coherent attention patterns across many local regions rather than produce a single global classification decision. Classification can partially average out local errors through the final global representation, whereas detection and segmentation depend more directly on stable token-to-token spatial relationships. From a deployment perspective, this makes bias-based methods particularly attractive for

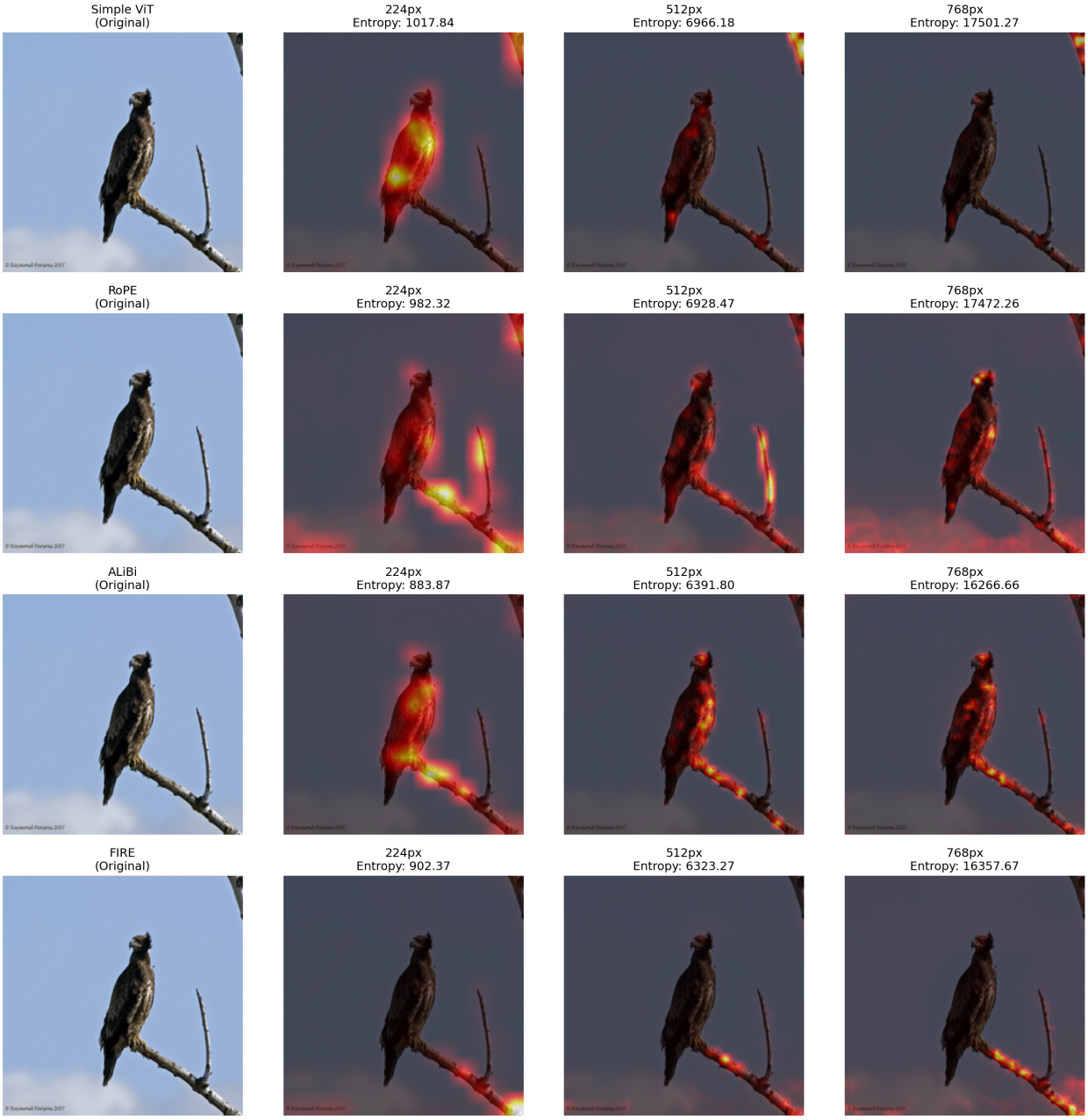


Figure 5: Attention visualization across resolutions. Each row shows a different positional encoding method; columns show the original image and attention heatmaps at increasing resolutions. Attention entropy (summed over all query-key pairs) quantifies attention diffusion. ALiBi maintains focused attention at 512 px (entropy 883.87) while other methods show 6–7× entropy increase. By 768 px, all methods exhibit high entropy, but ALiBi/FIRE retain visible concentration on the subject.

multi-scale vision systems: they may sacrifice some peak in-distribution accuracy, but they provide significantly more reliable behavior when the input resolution shifts at test time.

Overall, our findings support a clear practical recommendation. If the target application requires robustness across varying or unseen resolutions, additive bias-based positional encodings, especially FIRE and ALiBi, offer the strongest trade-off between accuracy and stability. If maximizing in-distribution accuracy is the primary objective and the evaluation resolution closely matches the training setup, RoPE remains highly

Table 5: Computational cost at varying resolutions (batch size 1). Memory in MB; throughput in images/second. Bold indicates best (lowest memory / highest throughput); underline indicates second best.

Model	Metric	224	384	512	768	1024
Simple ViT	Mem.	38	90	207	901	2755
	Thr.	**113**	**106**	**106**	**46**	**17**
RoPE ViT	Mem.	40	96	216	921	2791
	Thr.	80	83	80	36	13
YaRN ViT	Mem.	**35**	**48**	**73**	**195**	**491**
	Thr.	75	64	67	33	13
ALiBi ViT	Mem.	44	141	362	1690	5249
	Thr.	99	101	94	28	10
FIRE ViT	Mem.	45	165	453	2165	6765
	Thr.	107	105	84	17	6

competitive, but its extrapolation fragility should be carefully considered. YaRN provides a practical middle ground: by selectively rescaling RoPE frequencies at inference time without retraining, it substantially improves extrapolation robustness over raw RoPE while retaining much of RoPE's in-distribution strength.

5.5 Computational Cost

Table 5 reports peak GPU memory and inference throughput across resolutions. All methods have similar parameter counts (∼5.5M), but differ substantially in memory and compute overhead. Bias-based methods (ALiBi, FIRE) require storing the full $N \times N$ attention bias matrix, leading to higher memory at large resolutions: FIRE uses 6.8 GB at 1024 px versus 2.8 GB for Simple ViT. YaRN shows anomalously low memory due to implementation differences in frequency caching. Throughput follows inverse trends: Simple ViT achieves 113 img/s at 224 px while FIRE drops to 5.8 img/s at 1024 px.

These computational trade-offs should inform deployment decisions. While bias-based methods offer superior extrapolation, their 2–2.5× memory overhead at high resolutions may be prohibitive in resource-constrained settings. Conversely, if memory is limited but extrapolation robustness is required, YaRN provides a practical compromise with minimal overhead over RoPE.

6 Conclusion

We presented a systematic study of positional encoding strategies for Vision Transformers under resolution extrapolation, evaluating learned absolute embeddings Dosovitskiy et al. (2021), RoPE Su et al. (2024); Heo et al. (2024), ALiBi Press et al. (2022), YaRN Peng et al. (2024), and FIRE Li et al. (2024) across classification, detection, and segmentation. Our experiments reveal a clear dichotomy: attention bias-based methods (ALiBi, FIRE) substantially outperform rotation-based methods (RoPE, YaRN) under extrapolation, with FIRE retaining 63.1% Top-1 accuracy at 4.6× resolution compared to RoPE's 18.9%. Attention analysis confirms that this performance gap stems from attention pattern stability: bias-based methods maintain focused, semantically meaningful attention, while rotation-based methods exhibit attention collapse at extrapolated resolutions.

Practical Recommendations. For applications requiring resolution flexibility, we recommend FIRE or ALiBi as they provide the best extrapolation-accuracy trade-oaff. RoPE remains competitive for fixed-resolution deployment where in-distribution accuracy is prioritized. YaRN offers a practical middle ground when modifying a pre-trained RoPE model, providing substantial extrapolation gains through inference-time frequency scaling without retraining.

Limitations. Due to computational constraints, our study focuses on ViT-Tiny (∼5.7M parameters) trained on ImageNet-100. While the consistent trends across three diverse tasks suggest our findings generalize, validation on larger models (ViT-Base/Large) and full ImageNet-1K would strengthen these conclusions. We report single-seed results; however, the convergent patterns across classification, detection, and segmentation provide implicit robustness evidence. Additionally, the low absolute mAP values in detection reflect ViT-Tiny's limited capacity rather than fundamental limitations of the positional encoding methods.

Future Directions. Several avenues merit further investigation: (i) scaling analysis to determine whether extrapolation trends hold for larger ViT variants; (ii) learned frequency scaling for RoPE that adapts to 2D spatial structure rather than using fixed NTK-by-parts ramps; (iii) hybrid approaches combining the in-distribution strength of rotation-based methods with the extrapolation stability of bias-based methods; and (iv) extension to video transformers where temporal extrapolation introduces additional challenges. We hope this work provides a foundation for building resolution-robust vision architectures under practical compute constraints.

Broader Impact. Resolution-flexible Vision Transformers can reduce computational costs and carbon footprint by enabling training at lower resolutions while maintaining performance at deployment. This work contributes to more efficient vision systems, though practitioners should validate extrapolation behavior on their specific domains before deployment.

References

Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Better plain ViT baselines for ImageNet-1k. In *arXiv preprint arXiv:2205.01580*, 2022.

Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *CVPR*, 2023.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open MMLab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. In *EMNLP*, 2024.

Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *ICLR*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Zican Dong, Junyi Li, Xin Men, Wayne Xin Zhao, Bingbing Wang, Zhen Tian, Weipeng Chen, and Ji-Rong Wen. Exploring context window of large language models via decomposed positional vectors. In *NeurIPS*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Wei Han, Pan Zhou, and Shuicheng Yan. Stacked from one: Multi-scale self-injection for context window extension. *arXiv preprint arXiv:2603.04759*, 2026.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *ECCV*, 2024.

Xiangyu Hong, Jia Che, Biqing Jiang, Fandong Qi, and Yu Mo. On the token distance modeling ability of higher RoPE attention dimension. In *Findings of EMNLP*, 2024.

Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. Functional interpolation for relative positions improves long context transformers. In *ICLR*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

Xiaoran Liu, Yuerong Yin, Zhigeng Liu, and Fangming Huang. Beyond real: Imaginary extension of rotary position embeddings for long-context LLMs. *arXiv preprint arXiv:2512.07525*, 2025.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

MMSegmentation Contributors. MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *ICLR*, 2024.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *ICLR*, 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *NAACL*, 2018.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554*, 2022.

Qingyuan Tian, Zhijian Wenhong, Liaofeng Xiao, and Ru Wang. MrRoPE: Mixed-radix rotary position embedding. *arXiv preprint arXiv:2601.22181*, 2026.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Chaojun Xiao, Pengle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. InfLLM: Training-free long-context extrapolation for LLMs with an efficient context memory. In *NeurIPS*, 2024.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018.

Lei Yang, Shaoyang Xu, Jianxiang Peng, Shaolin Zhu, and Deyi Xiong. DCIS: Efficient length extrapolation of LLMs via divide-and-conquer scaling factor search. In *EMNLP*, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017.

Tongyao Zhu, Qian Liu, Wang Haonan, Shiji Chen, and Min Xiang. SkyLadder: Better and faster pretraining via context window scheduling. In *NeurIPS*, 2025.